# TODD WARCZAK, Ph.D.

Data Scientist

## CONTACT


twarczak@gmail.com
(206) 999-6478
Pleasanton, CA
toddwarczak.netlify.app
LinkedIn Github @TWarczak


## EDUCATION

**Ph.D.**

Molecular & Cellular Biology
Dartmouth College
September 2012 - August 2020

**B.S.**

Biology - University of Utah
September 2008 - May 2012

## SKILLS

**Coding**

R ★★★★, Python ★★★, SQL ★★★

**Cloud**

AWS EC2/S3, Docker, Snowflake

**Reports/Collaboration**

R-Shiny, Github, Quarto, Markdown, Jupyter, Jira

**Data Cleaning/Wrangling**

Tidyverse, Pandas, NumPy

**Custom/Automated Data Viz**

ggplot2, Plotly, reactable, Leaflet/ggmaps/sf (geospatial), modeltime (time-series)

**Machine Learning**

Tidymodels, SageMaker, scikit-learn, TensorFlow, Keras

**Time-Series Forecasting**

modeltime, lubridate, timetk

**General**

Probability & Statistics, NLP
Microsoft Excel/PowerP/Word,
Regression/Classification/Clustering

## WORK EXPERIENCE

### Data Scientist II

Bio-Rad Laboratories, January 2022 - March 2023 / Pleasanton, CA

- Developed "ddTrackR", a highly modularized R-Shiny application for team of scientists to upload experimental ddPCR results with metadata to database on AWS cloud. Expanded application for team to que/download data, build custom ggplots, and perform statistical calculations in browser, without code.
- Built ddPCR and NGS data analysis pipelines for 10+ scientists. Markdown reports delivered and/or Shiny dashboards hosted for team to visualize/explore results.
- Scrum master for team of 6 software developers building Bio-Rad proprietary tools in Python and Java for ddPCR QX600 machine. Utilized Jira Software for sprints.
- Hosted "Data Science" trainings every other week for 15+ scientists to develop skills in R and Python. Workshop topics included RStudio/VS Code setup, data wrangling with base-R/Tidyverse/Pandas/NumPy, statistics, mastering ggplots, exploratory data analysis, custom & dynamic {reactable} tables, ddPCR/NGS analysis, and more.

### Molecular Biologist

Dartmouth College, September 2012 - September 2020 / Hanover, NH

- Engineered novel genome-wide association study (GWAS) that identified genes controlling arsenic tolerance in plant roots. Utilized expert data cleaning, wrangling, and analytic skills to summarize findings from millions of observations across thousands of unique genomes.
- Determined plant gene AtNIP1;1 is the major genetic factor for tolerating arsenic in root cells and identified regions of interest on multiple chromosomes.
- Built lab RNA-seq pipeline for gene expression of 25000+ plant genes and wrote R scripts for gene clustering (PCA, hierarchical clustering), regression (GLMs/ANOVA), exploratory data analysis, and causal inference.

## SIDE PROJECTS OF NOTE

### SageMaker + RStudio to Predict Home Prices w/ Multi-class XGBoost; Explaining Model Behavior with Geospatial Plots and SHAP

- Explored Austin dataset to predict home price in Kaggle competition. (Blog, Github)
- Built static and interactive geospatial plots overlayed with feature data.
- Feature engineered high/low important words that associate w/ price using NLP.
- Trained/tuned/evaluated/deployed SageMaker Multi-class XGBoost on holdout data & submitted predicted binned price to Kaggle competition. Submission scored 0.8876 (mLogLoss), which would have placed 6th (out of 90 entries) in live competition.
- Modified {SHAPforxgboost} package to generate multi-class SHapley Additive exPlanations (SHAP) values/plots that explain how XGBoost model made predictions.

### Forecasting Daily Sales with {modeltime}

- Performed EDA and generated 3 month forecasts of daily sales for Kaggle dataset selling furniture, technology, and office supplies. (Blog, Github)
- Tested multiple {tidymodels} workflows with {modeltime} for time-series forecasting.
- Best individual models combined to generate single weighted ensemble forecast (Support Vector Machine/Neural Network-AR/Random Forest/Prophet-XGboost).